Instructional Science
https://doi.org/10.1007/s11251-020-09526-1

ORIGINAL RESEARCH

It matters how to recall – task differences in retrieval practice

Tino Endres[1] · **Lena Kranzdorf**[1] · **Vivien Schneider**[1] · **Alexander Renkl**[1]

Received: 7 January 2019 / Accepted: 26 October 2020

Abstract
The type of a recall task may substantially influence the effects of learning by retrieval practice. In a within-subject design, 54 university students studied two expository texts, followed by retrieval practice with either short-answer tasks (targeted retrieval) or a free-recall task (holistic retrieval). Concerning the direct effects of retrieval practice, short-answer tasks led to increased retention of directly retrieved targeted information from the learning contents, whereas free-recall tasks led to better retention of further information from the learning contents. Concerning indirect effects, short-answer tasks improved metacognitive calibration; free-recall tasks increased self-efficacy and situational interest. These findings confirm the assumption that the effects of retrieval practice depend on the type of recall task: short-answer tasks help us remember targeted information units and foster metacognitive calibration. Free-recall tasks help us remember a broader spectrum of information, and they foster motivational factors.

Keywords Retrieval practice · Task-specific effects · Spreading activation · Testing effect · Test-potentiated learning

Introduction

Retrieval practice is a learning activity that involves recalling information from memory. Retrieval practice enhances retention in contrast to restudying the material, a robust effect known as the testing effect. Meta-analytic reviews have shown that this effect is usually robust and reveals a medium to large effect on learning outcomes (e.g. Pan and Rickard 2018, Cohen's $d = 0.40$; Rowland 2014, between-subject designs Hedges' $g = 0.69$; within-subject designs Hedges' $g = 0.43$).

The effects of retrieval practice are one of the findings from cognitive psychology that is easily exploitable in different educational contexts, such as school settings (e. g. Butler and Roediger 2007; Pyc et al. 2014), university settings (Carpenter et al. 2016), or online learning (Davis et al. 2016; for recent reviews, see Adesope et al. 2017; Carpenter 2012; Dunlosky et al. 2013; Pan and Rickard 2018; Rowland 2014).

✉ Tino Endres
tino.endres@psychologie.uni-freiburg.de

[1] Educational and Developmental Psychologie, Albert-Ludwigs-University Freiburg, Engelbergerstraße 41, Freiburg, Germany

Within this huge body of research, different types of retrieval tasks are used when applying or analyzing the effects of retrieval practice (e.g. Chan and McDermott 2007; Karpicke and Aue 2015; Zaromb and Roediger 2010). However, there has been little research on which types of retrieval are best applied in educational contexts. A teacher may wonder if she should use more general tasks such as "What do you remember from the text about trees?" (holistic retrieval) or more specific recall tasks such as "What impact does the shape of leaves have on trees?" (targeted retrieval).

The current literature might suggest that the use of different types of retrieval tasks makes little difference (Rowland 2014). However, the present study provides empirical findings that the type of retrieval task does matter, and we demonstrate under which circumstances which task type is perhaps best. Such findings help instructors choose the best-suited type of retrieval task based on their educational goals. Moreover, they provide insights into the mechanisms involved in retrieval practice.

More specifically, we analyzed the effects of retrieval on different educationally-relevant aspects. We investigated the *direct effects* of retrieval on learning different types of information and the *indirect effects* on factors that likely impact further learning such as metacognitive calibration (correctness of metacognition) and motivation (here: situational interest and self-efficacy, Arnold and McDermott 2013).

Direct effects of retrieval practice

Regarding direct effects of retrieval practice, there is ample evidence for differences when using different retrieval-practice tasks. However, the vast majority of studies focus on comparing recognition-oriented tasks to recall-oriented tasks; they have usually found learning to be optimal when the final learning assessment employs the same kind of tasks as the test-for-learning (recognition—recognition, recall—recall; e.g. Duchastel and Nungester 1982). When the learning goal is recognition-oriented, learners will profit most from recognition-oriented tasks during retrieval practice. When the recall of materials is the goal, learners should profit most from recall tasks during retrieval practice. From a theoretical point of view, these findings are in line with the transfer-appropriate processing view of retrieval practice (Morris et al. 1977; Veltre et al. 2016).

We wondered whether there are not more task-specific differences. We were particularly interested in the recall of learning contents, as recall is necessary in most application situations. Theoretically speaking, even various types of recall tasks such as those that are more or less specific (i.e. short-answer tasks or free-recall tasks) should trigger differences in the learning process and thereby lead to different outcomes.

There are multiple theories about how retrieval-practice effects occur (for an overview see Karpicke et al. 2014). One of the most prominent is the elaborative retrieval theory (e.g. Carpenter 2009). From this theory's perspective, different retrieval practice tasks should make a difference in retention even when both tasks are recall tasks. The elaborative retrieval theory attributes the benefit of recalling learning materials to two processes: Firstly, the recall of a specific piece of knowledge activates the memory traces of that specific piece of knowledge. Activating this memory trace strengthens the connections between the concepts related to that specific memory trace. These strengthened connections ultimately lead to better retrieval of that memory trace. Secondly, the mental effort invested in recall—induced by search processes in memory—spreads activation to other pieces of knowledge that are connected to the retrieved piece of knowledge; the memory traces of those connected knowledge pieces and their association with the targeted piece of

knowledge are thereby strengthened. This process of spreading activation eventually also leads to better retention of the targeted knowledge pieces and of the connected knowledge pieces. Such spreading activation can even lead to better learning performance on knowledge pieces that were not recalled during retrieval practice (retrieval-induced facilitation; Chan et al. 2006).

When learning by retrieval practice, the two mechanisms of strengthening and spreading activation should lead to a particular activation pattern that depends on the recall task's specificity. If a recall task is specific and requires merely a targeted piece of information, for example, from a text (e.g. short-answer task), this task should lead to much greater activation of that specific, targeted information piece, as compared to the activation of other information pieces from the text. Such targeted retrieval should consequently lead to better retention of the content sections that were targeted by specific recall tasks. In comparison, if a recall task is unspecific and asks for a variety of information pieces (e.g. free-recall task), there should be less of an activation advantage for one specific piece of information; a variety of information pieces may be activated (Anderson and Reder 1999). In other words, the degree of activation generally available must probably be shared between more information pieces.

Few studies have investigated the effects of different types of recall as retrieval practice tasks. Glover (1989), for example, compared restudy, cued recall, free-recall, and recognition, identifying differences between the tasks and free-recall to be most beneficial for learning. However, Glover compared only free-recall with all other conditions combined. He did not specifically compare short-answer tasks and free-recall tasks. In a more recent study, comparisons of short-answer and free-recall retrieval-practice tasks revealed no task-specific differences (Endres and Renkl 2015). A more structured recall process that divided answers by specific prompts revealed no significant differences compared to free-recall retrieval practice (Smith et al. 2016). Taken together, these studies identified no differences in learning between open recall tasks and more specific recall tasks.

Few studies have directly compared different types of recall tasks (note, as mentioned above, most other studies investigated differences between recognition-oriented tasks and recall tasks, e.g., Kang et al. 2007; Rawson and Zamary 2019; Smith and Karpicke 2014). A meta-analysis can, however, also compare the effects sizes of studies using different types of recall tasks. Rowland's meta-analysis (2014) showed that differences in the specificity of tasks do not lead to significant differences as long as the recall rate during retrieval practice is high. The effect sizes of both task types were similar (the short-answer referred to as cued recall in the meta-analysis, Hedges' $g = 0.72$; free recall, Hedges' $g = 0.81$).

One drawback of the aforementioned meta-analysis and of previous studies is that the learning was assessed in a rather coarse manner. Usually, only general learning outcomes were tested; different information pieces from the text (e.g. whether an information piece was targeted in the recall task or not) were not differentiated. Such coarse assessments of learning outcomes do not enable us to test more detailed theoretical assumptions of the effect of different types of recall tasks, although such differences might still be relevant when designing learning arrangements including retrieval practice.

Our study differentiated various content sections when assessing learning outcomes to gain both practically and theoretically relevant insights. If activation is the relevant mechanism predicting retrieval practice effects, then there should be clear differences when either free-recall (holistic retrieval) or short-answer tasks (targeted retrieval) are used. We formulated the following hypotheses:

Targeted retrieval hypothesis: Targeted retrieval via specific short-answer tasks improves the learning of the targeted sub-content more than holistic retrieval by free-recall tasks.

Holistic retrieval hypothesis: Holistic retrieval induced by an unspecific free-recall task improves the learning of various sub-contents more than targeted retrieval by specific short-answer tasks.

Indirect effects of retrieval practice

Different recall tasks may not just lead to different learning outcomes—they can influence other factors relevant to (future) learning. Influences on such factors are referred to as indirect effects of retrieval practice (Roediger et al. 2011). These effects are especially relevant when implementing retrieval practice in practice settings, particularly in self-regulated learning arrangements. In such settings, students can decide whether to continue studying, which content to restudy, and how much effort to invest. Although the direct effects of retrieval-based learning can have a substantial impact, the benefits of retrieval-based learning as a whole can only be fully understood when potential remedial learning activities after first recall are also taken into account. Hence, for practical purposes, it is particularly important to know about both the direct and indirect effects of various types of recall tasks.

A key issue regarding indirect effects is the kind of information the learners obtain about their previous success on a recall task. This information influences how accurate the learners monitor their own state of knowledge and knowledge gaps. Such meta-knowledge is important, as it influences learners in their decision about starting re-learning and investing more or less mental effort in such relearning. Learners' efforts to close their perceived knowledge gaps are termed regulation. It is in particular the learners' perception about the amount of knowledge they have already acquired that determines whether they will invest more or less time and effort in restudy, that is, the remedial learning of specific contents (Nelson and Narens 1990; Nückles et al. 2009).

Learners' monitoring of knowledge states and gaps can be influenced by many information sources. One classic source is feedback from the teacher who evaluated the learner's performance on a test or a recall task. Beyond this direct feedback, learners can also monitor themselves, for example, by relying on inherent feedback when working on (recall) tasks (Arnold and McDermott 2013). The fluency of recall and perception of knowledge gaps are crucial factors in this context (Koriat 2012).

Let us reconsider specific and unspecific recall tasks. If a task demands a holistic activation of contents, for example, in a free-recall task, the inherent feedback the learner receives about knowledge gaps is relatively sparse. Although learners may miss some important information, they might not notice, because they can write about other aspects in response to the recall task. The process of recall might be fluent, and the answer provided may appear correct and coherent. As a result, learners get the impression that their answers were good even though significant pieces of knowledge might be missing (Koriat 2012).

If a task targets a very specific answer, for example, a short-answer task, the inherent feedback learners get while monitoring is much higher. For example, there may be several short-answer tasks on the same content section as for a free-recall task. These short-answer tasks require specific, detailed answers, and learners might become aware of their inability to provide such answers. They thus come to realize that they have knowledge gaps. Moreover, the process of remembering while working on a short-answer task might not be fluent, and the subjective probability of having provided a correct answer might be relatively low.

Overall, the inherent feedback the learners obtain from specific tasks is higher, and it provides them with a reasonable idea of their "objective" knowledge state (Koriat 2012).

The more feedback a task provides to learners, the more accurate their meta-knowledge should be, that is, they are better able to calibrate (Alexander 2013). Short-answer tasks provide more intrinsic feedback than free-recall tasks. Hence, we formulated the following hypothesis:

Calibration hypothesis: Targeted retrieval via specific short-answer tasks increases calibration more than holistic retrieval via an unspecific free-recall task.

We also assume that differences in the success learners perceive while recalling can influence motivation. Self-efficacy and situational interest may be influenced especially by the type of recall task. Self-efficacy builds mainly on our previous perceptions doing a certain activity (Bandura 1997). When learners have perceived success, they also expect to succeed in the future with similar tasks (Schunk 1985). Returning to the task-specific effects in feedback from the paragraphs above: the inherent feedback in recall tasks might also exert influence on perceived success. Hence, free-recall tasks—with their low level of intrinsic feedback and relatively fluent recall as compared to short-answer tasks—should lead to a stronger perception of success. After perceiving such success, learners should have higher self-efficacy expectations after a free-recall task.

Situational interest is also affected by perceived success. As with self-efficacy, learners are more interested in areas in which they perform well (Hidi and Renninger 2006). The concepts of self-efficacy and situational interest are closely interconnected, according to some theories (e.g. Schiefele 1990). Hence, free-recall may lead to a stronger experience of success—compared to short-answer tasks – and thus should also trigger greater situational interest. On the other hand, there are perspectives predicting greater situational interest after a short-answer task. The knowledge-deprivation hypothesis predicts that a perceived lack of knowledge leads to higher situational interest. According to the assumption that short-answer tasks entail a high level of intrinsic feedback on knowledge gaps, such tasks should trigger stronger situational interest (Rotgans and Schmidt 2011, 2014).

We formulated the following hypothesis:

Self-efficacy hypothesis: Holistic retrieval via unspecific free-recall increases self-efficacy more than targeted retrieval via specific short-answer tasks.

Situational interest hypothesis: The type of recall task influences situational interest (two-sided hypothesis).

Method

Sample and design

We employed G*Power (Faul et al. 2007) as statistical power analysis software to estimate the minimum sample size for our within-subjects design. The software estimated that 34 participants would be needed to detect a statistically significant difference for the assumed medium effect size (effect size Cohen's $d = 0.5$, α-level $p = 0.05$, power 80%). Our study enrolled even more participants, thus enabling us to demonstrate reliable effects.

The 54 university students (age: $M = 22.50$, $SD = 5.44$) participating in this study were majoring in different subjects. Participants were given course credit for participation. All were aware they were taking part in research. The experimenter informed each participant about the possibility of quitting the experiment with no repercussions or drawbacks at any

time. All participants provided informed consent and permitted us to use their collected data anonymously for publication.

We applied a within-subjects design with the factor "type of recall task". The factor consisted of a holistic retrieval condition by unspecific free-recall and a targeted retrieval condition by specific short-answer. As dependent variables, we assessed learning outcomes determined by a posttest including free-recall tasks and short-answer tasks (direct effects) as well as metacognition, self-efficacy, and situational interest (indirect effects).

Materials

Texts

We drafted two texts (Text 1: 2427 words; Text 2: 2607 words) dealing with two different contents (coffee and sugar). Each text consisted of four sections similar in length, structure, and complexity. To understand a specific paragraph, learners did not need to understand the previous paragraphs. There were no references between the paragraphs. Each section contained three important pieces of interconnected information (see Appendix A). This separation into textual sections enabled us to detect fine-grained differences in learning outcomes.

The texts were assessed for their readability, intelligibility, and reading time (M = 20.6 min, *SD* = 4.07, max. 25 min) in a pilot study. Twelve students read both texts and rated the readability and intelligibility of both texts on four rating items (see Appendix B). The texts did not differ in any of the four items (all $ps > 0.7$). After reading and rating, the texts were returned to the students, and they were asked to mark those paragraphs that needed improvement. We improved the marked paragraphs. The original texts were in German (see Appendix C for English translations, please note, due to translation some characteristics of the text might have changed. If you are interested in the original materials, please contact us).

Recall tasks

We constructed short-answer tasks and free-recall tasks. The short-answer tasks were used to induce the activation of specific targeted information. We constructed 12 specific short-answer tasks for every text. Each text had three questions targeting a specific section in the text. Each question assessed one aspect required to understand the text (e.g. "Which characteristics of a coffee plant should be considered for agricultural purposes?", see Appendix D). For the short-answer tasks, participants were asked to give an answer consisting of 2 to 4 sentences. Text under each answer box prompted learners to provide an answer of comparable length by using a counter for used characters (200–400 characters). The characters were counted while they typed. The count was displayed under the textbox. The color of the count turned from red to green once 200 characters had been reached and turned red again once 400 characters had been reached. Nevertheless, the participants could provide shorter or longer answers.

We used all short-answer tasks in the posttest. The provision of tasks in the learning phase depended on the condition. We scored each answer to a short-answer task on a scale ranging from 0 to 3 (with possible partial credit of 0.5). The maximum score was 36 points per text. The scales' consistency was acceptable for complex learning items (Schmitt 1996; sugar Cronbach's $\alpha = 0.65$; coffee Cronbach's $\alpha = 0.59$). Two individual scorers rated the

learners' answers. We double-coded 37% of the data (20 participants). Interrater reliability was high (ICC = 0.95).

Free-recall tasks were used to induce the holistic activation of different learning contents. The free-recall task asked learners to recall all the aspects in the text they could remember (e. g. "What do you remember about the coffee text?"). For the free-recall task, participants were asked to provide an answer consisting of 6 to 12 sentences, a number of sentences that is three times more than a single short-answer task required. A text under each answer box prompted learners to provide an answer of comparable length by using a counter (600–1200 characters). The color of the count turned from red to green once 600 characters had been reached and turned red again once 1200 characters had been reached. The participants could nevertheless provide shorter or longer answers.

The same tasks functioned to assess learners' knowledge in the posttest. The maximum score was, as with the short-answer task, 36 points per text. The learners' answers were scored by two individual raters. We double coded 20 participants which corresponds 37% of the data. Interrater reliability was high (ICC = 0.92).

Mental effort

To assess subjective mental effort, we asked participants after each recall task type how much effort they had invested in answering the task(s) (Pass 1992; Sweller et al. 2011). Participants indicated their mental effort using a scroll-bar (range: 1 [= low] to 9 [= high]).

Calibration

To assess the meta-knowledge for calculating calibration, we used two judgment-of-learning items (JOLs). Once the participants had worked on each text's recall tasks, we asked them to indicate how correct they thought their answers were and how probable it was that they would remember their answers next week. Participants indicated their meta-knowledge using a scroll bar (range: 0% [= low] to 100% [= high]). As both scores correlated substantially, we aggregated them into one score (Cronbach's $\alpha = 0.741$).

For the meta-knowledge score, we considered the aggregated JOLs as a predicted score for the posttest (percentage of maximum posttest score). We compared this predicted score to the (objective) knowledge score in the delayed assessment and calculated a discrepancy score using the absolute value of discrepancy (Schraw 2009).

Self-efficacy

We assessed self-efficacy with five self-rating items after working on each text's recall tasks. The five items assessed learners' anticipated ability to perform similar tasks in different situational and social situations (e.g., could you explain the (the topic) to a friend, could you answer a similar question(s) in a potential exam). We followed Bandura's guidelines (2006) for constructing self-efficacy scale in how we formulated our items. Participants indicated their self-efficacy using a scroll bar (range: 0% [= low] to 100% [= high]; Cronbach's $\alpha = 0.904$).

Situational interest

We assessed situational interest by relying on three self-rating items after working on each text's recall tasks (Schiefele 1990). The items asked whether the participants felt that the content presented was interesting, entertaining, or boring. Participants stated their situational interest using a five-point rating scale (range: 1 [=low] to 5 [=high]). The answers to the boredom item were reversed (Cronbach's $\alpha = 0.827$).

Procedure

Our experiment consisted of two computer-based sessions that entailed four phases: the learning phase, intervention phase (recall tasks: targeted retrieval by short-answer or holistic retrieval by free-recall), immediate assessment phase (metacognition, self-efficacy, and situational interest) in the first session, and the delayed-assessment phase in the second session (posttest on learning outcomes in short-answer and free recall posttest; delay: 7 days).

In the first session, participants started with the learning phase in which they studied two expository texts, each text introduced with the same sentence: "Please read the following texts carefully. There will be a knowledge test." (translated from German by the first author). Once both texts had been read, the intervention phase began. Targeted activation and holistic activation were implemented via different recall tasks (targeted retrieval by short-answer tasks and holistic retrieval by free-recall tasks). In our within design, the different task types were randomly assigned to one of two texts to control for (non-expected) text effects. The sequences of the texts and tasks were randomized. No feedback was provided after the tasks. After each task type, we assessed their mental effort to control for (unexpected) task differences. Furthermore, we assessed situational interest, metacognition, and self-efficacy after working on each text's recall tasks.

After a one-week delay, learners returned for the second session in which we assessed learning outcomes via a free-recall posttest and short-answer posttest for both texts. Participants had to answer the free-recall items in both texts first. After answering the free-recall items, participants had to answer twelve short-answer items in each text. These 24 items included three items that the participants had answered before.

The delayed assessment of learning outcomes is standard in investigations of retrieval practice and testing effects (see Rowland 2014). The tasks' order with regard to the two texts was randomized. The order of the task types was always the free-recall first, followed by the short-answer.

Results

We used an alpha level of 0.05 and relied on two-sided testing for all statistical analyses. We determined η^2 as an effect size index. The values 0.01, 0.06, and 0.14 were considered as small, medium, and large effect sizes, respectively. For certain analyses, we employed Bayesian analysis (Bayes Factor; BF) to confirm null hypotheses. We used a predefined, uninformative ZJS prior with a multivariate Cauchy distribution to (Rouder et al. 2012). This conservative prior makes no predictions about the effect to be found. We used this conservative prior because we wanted to ensure that evidence for a null effect would be conclusive. With this prior, the probability that the Bayes Factor reveals insufficient

evidence is higher than a false confirmation of the null hypothesis. We interpreted the Bayes Factor as did Wetzels et al. (2011).

Prior analyses

As a basis for analyzing task-specificity effects, we examined whether the mental effort would be similar in both conditions. There were no differences in mental effort ratings between the two conditions, $F(1, 54) = 0.286$, $p = 0.596$. In addition to our frequentist analysis, we conducted a Bayesian repeated measures ANOVA. The Bayes analysis revealed substantial evidence favoring the null hypothesis ($BF_{01} = 4.313$). This Bayes factor can be interpreted as follows: it is 4.313 times more likely that there is no difference between mental effort ratings between experimental conditions. This expected result is crucial for testing effects resulting from differential activation patterns induced by different task types, and it enabled us to interpret our data without assuming any mental effort difference between task types (Endres and Renkl 2015; Pyc and Rawson 2009).

Direct effects

To assess the direct effects, we checked for differences in the one-week-delayed assessment phase. Table 1 provides the descriptive statistics of performance variables in both conditions. We had no participant unable to retrieve any information, and observed no differences between targeted retrieval and holistic retrieval conditions on overall performance (short-answer posttest: F [1, 54] = 0.458, $p = 0.502$; free-recall posttest: F [1, 54] = 0.892, $p = 0.349$). In addition to frequentist analysis, we conducted a Bayesian repeated measures ANOVA. The Bayes analysis revealed substantial evidence favoring the null hypothesis (short-answer posttest: $BF_{01} = 4.091$, free-recall posttest: $BF_{01} = 3.364$). This Bayes factor can be interpreted as follows: It is 4.091 or 3.364 more likely that there is no difference between learners' performance between experimental conditions. On the overall performance level, our study revealed no task-specific performance differences (Bayesian analysis), in line with Rowland's (2014) recent meta-analyses.

On the more detailed level of content sections, however, we identified several differences between task types. We analyzed those according to our hypotheses:

Table 1 Means and standard deviations (in brackets) of performance variables in both conditions

	Posttest format	Targeted retrieval condition	Holistic retrieval condition
Overall	Short-answer	11.63 (4.84)	11.28 (3.80)
	Free-recall	9.24 (4.41)	9.81 (4.37)
Targeted information	Short-answer	3.63 (1.76)	2.82 (0.95)
	Free-recall	2.93 (1.69)	2.45 (1.09)
Non-targeted information	Short-answer	2.67 (1.26)	
	Free-recall	2.10 (1.17)	

Targeted retrieval hypothesis

To test the targeted retrieval hypothesis ("Targeted retrieval via specific short-answer tasks improves the learning of the targeted sub-content more than holistic retrieval by free-recall tasks"), we compared the two conditions' learning outcomes with respect to knowledge pieces specifically addressed in the short-answer tasks. We expected that the targeted retrieval by specific recall (in contrast to holistic retrieval via free-recall tasks) would improve the performance in the targeted information section in both posttest types.

Targeted retrieval improved the learning outcomes of targeted content sections more than holistic retrieval in both posttest types (short-answer posttest: targeted retrieval in targeted retrieval condition, $M = 3.63$, $SD = 1.76$, targeted retrieval in holistic retrieval condition, $M = 2.82$, $SD = 0.95$, F [1, 53] $= 12.868$, $p < 0.001$, $\eta^2 = 0.195$; free-recall posttest: targeted retrieval in targeted retrieval condition, $M = 2.93$, $SD = 1.69$, targeted retrieval in holistic retrieval condition, $M = 2.45$, $SD = 1.09$, F [1, 53] $= 4.031$, $p = 0.04980$, $\eta^2 = 0.071$; see Table 1). We also observed this difference when aggregating both posttest types (F [1, 52] $= 6.644$, $p = 0.003$, $\eta^2 = 0.204$). The interaction between task type for practice and task type in the posttest was not statistically significant (F [1, 53] $= 1.621$, $p = 0.208$, $\eta^2 = 0.030$). Hence, we found no evidence for a transfer-appropriate processing effect.

In summary, we confirmed the *targeted retrieval hypothesis*. In both posttest types, the specifically recalled content sections were recalled better than the not specifically recalled ones. As the amount of mental effort and overall posttest performance did not differ between conditions, this difference is likely due to diverse activation patterns. The specific activation of one section strengthened this section more than did the activation of multiple sections.

Holistic retrieval hypothesis

To test our unspecific recall hypothesis ("Holistic retrieval induced by an unspecific free-recall task improves the learning of various sub-contents more than targeted retrieval by specific short-answer tasks"), we compared the learning outcomes between conditions with respect to knowledge pieces non-targeted via the short-answer tasks. We expected that the holistic activation via the free-recall task in contrast to non-targeted information in the short-answer tasks would improve performance in both posttest types.

Holistic retrieval improved learning outcomes of several sections more than target retrieval did (free-recall posttest: targeted retrieval in holistic retrieval condition, $M = 2.45$, $SD = 1.09$; non-targeted information in targeted retrieval condition, $M = 2.10$, $SD = 1.17$, see Table 1). This effect was, however, only apparent in the free-recall posttest types (F [1, 53] $= 4.841$, $p = 0.032$, $\eta^2 = 0.084$). It was not revealed in the short-answer posttest types (short-answer posttest: targeted retrieval in holistic retrieval condition, $M = 2.82$, $SD = 0.95$; non-targeted information in targeted retrieval condition, $M = 2.67$, $SD = 1.26$; F [1, 53] $= 1.289$, $p = 0.261$, $\eta^2 = 0.024$, overall: F [1, 52] $= 2.381$, $p = 0.102$, $\eta^2 = 0.084$). Following up this insignificant finding, we additionally conducted a Bayesian repeated measures ANOVA to obtain evidence for the null hypothesis. The Bayes analysis yielded no evidence favoring any hypothesis ($BF_{10} = 0.359$). The interaction between type of practice task and posttest task type was not statistically significant (F [1, 53] $= 1.655$, $p = 0.204$, $\eta^2 = 0.030$). Hence, there was no transfer-appropriate processing effect.

In summary, we confirmed the *holistic retrieval hypothesis* in the free-recall posttest. We could not confirm this hypothesis in the short-answer posttest.

Indirect effects

Calibration hypothesis

To test our calibration hypothesis ("Targeted retrieval via specific short-answer tasks increases calibration more than does holistic retrieval via an unspecific free-recall task") we compared the discrepancy score between the targeted-retrieval condition and holistic-retrieval conditions. Table 2 provides descriptive statistics of the indirect-effects variables in both conditions. Holistic retrieval leads to higher JOLs ratings than targeted retrieval (F [1, 53] = 11.470, $p = 0.001$, $\eta^2 = 0.179$). We calculated a difference from predicted posttest scores by using the JOLs of the learners and actual posttest scores. Targeted retrieval improved the accuracy of JOLs more than holistic retrieval (short-answer posttest: F [1, 53] = 11.470, $p = 0.001$, $\eta^2 = 0.178$; free-recall posttest: F [1, 53] = 11.470, $p = 0.001$, $\eta^2 = 0.178$; overall: F [1, 53] = 11.470, $p = 0.001$, $\eta^2 = 0.178$).

In summary, we confirmed the calibration hypothesis. Learners' calibration was more accurate after short-answer tasks than after a free-recall task.

Self-efficacy hypothesis

To test our self-efficacy hypothesis ("Holistic retrieval via unspecific free-recall increases self-efficacy more than targeted retrieval via specific short-answer tasks") we compared the aggregated self-ratings in the targeted retrieval and holistic retrieval conditions. Holistic recall raised self-efficacy more than targeted retrieval (F [1, 53] = 4.602, p = 0.037, $\eta^2 = 0.080$). To better understand the mechanisms underlying this effect, we conducted a within mediation analysis. We tested whether the effect of task type on self-efficacy was mediated by the participants' performance. We applied the MEMORE tool to calculate within mediations (Montoya and Hayes 2017). This model revealed indirect effects of condition on self-efficacy via a participant's performance, $B = 31.58$, 95% CI [9.81, 62.55].The insignificant c' path indicates a complete mediation (i.e. the effect is entirely attributable to the mediation; see Fig. 1).

In summary, we confirmed our self-efficacy hypothesis. Learners' self-efficacy was significantly higher in the holistic retrieval condition than the targeted retrieval condition. More intensely perceived success predicted greater self-efficacy after recall.

Table 2 Means and standard deviations (in brackets) of secondary effect variables in both conditions

	Posttest format	Targeted retrieval condition	Holistic retrieval condition
JOLs		43.54 (20.95)	54.42 (19.56)
Calibration	Short-answer	−0.75 (5.23)	1.98 (6.32)
	Free-recall	−0.39 (5.71)	2.33 (5.31)
Self-efficacy		48.12 (20.36)	54.48 (19.17)
Situational interest		3.36 (0.92)	3.83 (0.80)

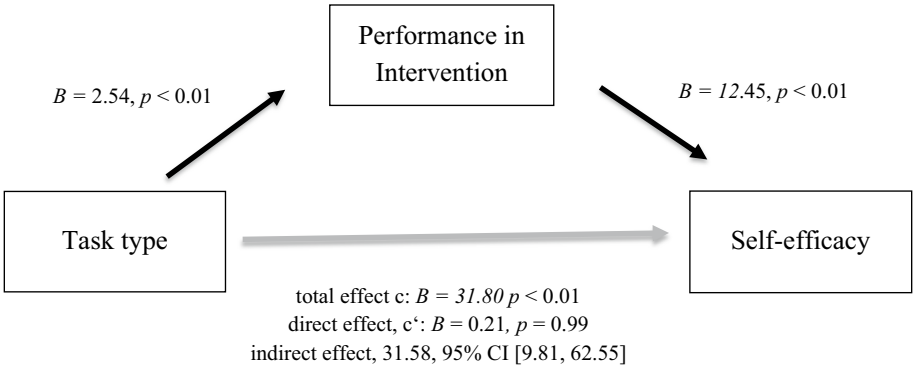


Fig. 1 Effect of task specificity on self-efficacy via participants' performance. All Bs represent unstandardized regression coefficients obtained through bootstrapping using 5000 resamples. The range in brackets represents the 95% CI of the indirect effect

Situational interest hypothesis

To test our situational interest hypothesis ("The type of recall task influences situational interest") we compared interest in the topics associated with the targeted retrieval condition and holistic retrieval condition, respectively. Holistic retrieval increased situational interest more than targeted retrieval (F [1, 53] = 13.928, $p < 0.001$, $\eta^2 = 0.208$). We conducted a regression analysis to better understand the mechanisms underlying this effect. We see two possibilities in our theoretical argumentation: If the parallel development of perceived success and situational interest support each other, a more intense experience of success would lead to deeper situational interest. The differences in recall in the test-for-learning should predict the situational interest differences with a positive b weight. A negative b weight would support the thirst-for-knowledge theory; in this case, the experience of knowledge gaps (negative experience of success) leads to greater situational interest. Specifically, we conducted a regression analysis using individual recall differences as a factor of "perceived success" to predict situational interest. Test-for-learning performance differences significantly predicted situational interest differences, $b = 0.33$, t (52) = 2.56, $p = 0.014$; the better the performance, the greater the situational interest. The test-for-learning performance also explained a significant proportion of the variance in situational interest, $R^2 = 0.11$, F (1, 52) = 6.53, $p = 0.014$. We also conducted a within mediation analysis, testing whether the effect of task type on situational interest was mediated by the participants' performance (MEMORE, Montoya and Hayes 2017). The model revealed an indirect effect of condition on situational interest via the participants' performance, $B = 0.83$, 95% CI [0.20, 1.77]. The insignificant c' path indicates a complete mediation (i.e., the effect is entirely attributable to the mediation; see Fig. 2).

In summary, the task type did in fact influence situational interest. Learners exhibited deeper situational interest after a free-recall task. A stronger experience of success predicted greater situational interest. In conjunction with low inherent feedback, such enhanced situational interest after having performed a task successfully supports the theory of the parallel development of experiencing success and situational interest (e.g. Hidi and Renninger 2006).

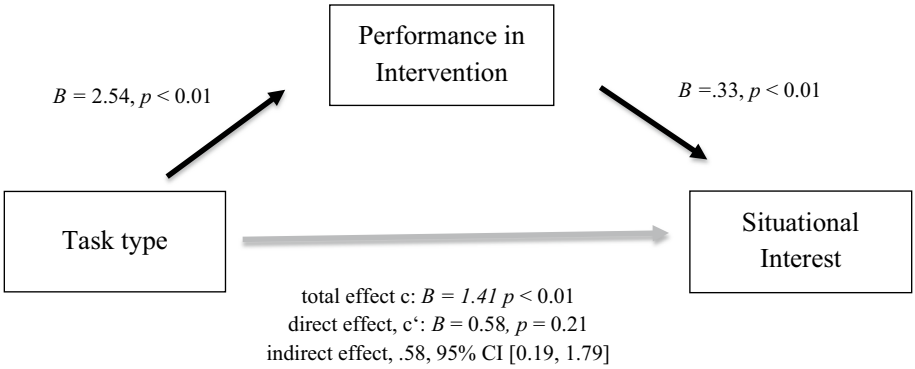


Fig. 2 Effect of task specificity on situational interest via participants performance. All Bs represent unstandardized regression coefficients obtained through bootstrapping using 5,000 resamples. The range in brackets represents the 95% CI of the indirect effect

Discussion

Our study makes the following contributions to the available literature: (1) The present findings confirm the assumption that task type matters when employing recall tasks for retrieval practice—a novel finding in this research field. Our study's setup has enabled us to demonstrate such novel evidence on the effects of specific and unspecific recall tasks. (2) We did not detect a mere better-or-worse pattern of results. Instead, our findings highlight the relevance of educational goals when implementing retrieval practice. Previous studies on retrieval practice barely addressed different educational goals. (3) Regarding the direct effects of retrieval practice, targeted retrieval (via specific short-answer tasks) led to greater retention of targeted information from the learning contents; holistic retrieval (via unspecific free-recall tasks) led to better retention of further information from the learning contents. (4) Regarding indirect effects, targeted retrieval (via specific short-answer tasks) improved metacognitive calibration; holistic retrieval (via unspecific free-recall tasks) increased self-efficacy and situational interest. Note that a strength of the present study is that we investigated both direct and indirect effects in a differentiated manner within a single study, thus enabling insights into the interdependency of direct and indirect task-type effects.

Theoretical implications

Regarding direct effects, our findings provide evidence of the mechanisms involved in retrieval-practice effects. Especially the process of spreading activation seems to play a role in our semantically related learning materials. Although our hypotheses have not yet been supported by empirical studies or meta-analyses (see also discussion, Rowland 2014), our findings suggest that there are systematic differences between tasks when the appropriate analytic grain size is chosen, that is, when learning contents and outcomes are analyzed in greater detail.

In our study, the distinctive activation patterns of knowledge we assumed (specific and unspecific) led to specific learning effects. We suggest that this specific learning effect is attributable to spreading activation as a relevant mechanism in retrieval practice. As

long as the topics contain meaningful cues (a factor we can assume in complex learning), spreading activation seems to have, at the very least, greater explanatory value than the transfer-appropriate processing perspective when considering differences between various types of recall task.

Our direct effect hypotheses were derived from the elaborative retrieval theory (Carpenter 2009). Spreading activation plays a crucial role in this theory. Although we did not design our study to (dis-) confirm theories, our findings and the current literature suggest that the transfer-appropriate processing perspective applies when different memory processes such as recognition and recall are in the foreground; however, theories including spreading activation possess greater explanatory power for explaining effects of different recall tasks. Follow-up studies should test this assumption.

Regarding indirect effects, it is the situational-interest effect in particular that is theoretically relevant. In our introduction, we presented two perspectives of how performance may influence situational interest (high performance may enhance or lower interest). Our subsequent mediation analysis indicated that a better-perceived performance leads to higher interest (see Fig. 2). This finding does not support the knowledge-deprivation hypothesis. However, there are major differences between our study and those supporting the knowledge-deprivation hypothesis (Rotgans and Schmidt 2014, 2017). In our case, the learners read both texts relevant to a later task performance. In studies on the knowledge-deprivation hypothesis, the topics addressed in the tasks were completely new to one group of learners. In conclusion the difference in knowledge-deprivation was greater in these studies. A second difference is the type of task carried out in knowledge-deprivation studies and in our (and other) retrieval-practice studies. In knowledge-deprivation studies (e.g. Rotgans and Schmidt 2014), the tasks were more like challenging puzzles instead of recall tasks. Students were given background information (a short introductory text) before they had to make a guess or work out a solution. Students in those studies were probably interested in whether their solution was right. The knowledge-deprivation hypothesis may have explanatory value for those task types, but less so for recall tasks.

Another interpretation of our effect of task specificity on situational interest may be in line with the self-determination theory. A free recall task could also lead to higher situational interest because participants can choose the topic they would like to recall. Such autonomy should trigger stronger situational interest (Ryan and Deci 2000), an assumption already validated (Linnenbrink-Garcia et al. 2013). Although this assumption is theoretically plausible, our study provides no evidence supporting it. Our mediation analysis revealed that the task effect on situational interest was fully mediated by performance. However, as we cannot entirely exclude a self-determination interpretation of the effect on interest. Perceived autonomy while working on different recall tasks should be investigated in future studies (cf. Vansteenkiste et al. 2004).

Practical implications

In this study, we learned that different recall tasks actually make a difference. We found that using specific tasks fosters the retention of those particular targeted content sections that had been included in these questions. On the other hand, unspecific tasks foster the learning of content sections that were not directly retrieved. Hence, when teachers decide on a task type for retrieval practice, they should also take the nature of the learning contents into account. If the contents to-be-learned include a few, very central content units, it would be preferable to use specific tasks such as short-answer tasks. There may be a

few specific concepts illustrated by several examples that facilitate understanding, but they themselves are not relevant. A text about three crucial factors of climate change, which also includes illustrations for better understanding, could serve as a corresponding example.

If the contents to-be-learned include a wider range of important information, it would be preferable to use unspecific tasks such as free-recall. For example, in a text on the Gulf Stream and different factors affecting it, there are many different and important points worth remembering.

If the contents to-be-learned include some very important information, but overall understanding also depends on learning a broad set of information, a teacher might apply both formats. For example, students might first work on free-recall tasks to deepen their overall understanding. Then, they could answer short-answer questions to consolidate the most important key information. This mix of specific and unspecific recall tasks may explain why mixed-format tests were associated with higher weighted mean effect sizes in a recent meta-analysis (Adesope et al. 2017).

With respect to indirect effects, our findings suggest that more specific task types (e.g. short-answer tasks) help learners become aware of their knowledge gaps, leading to more accurate metacognition. Free-recall tasks have the advantage of increasing self-efficacy and situational interest. Again, as in the case of direct learning outcomes, teachers should select a specific task type according to the learning unit's most important educational goals. If teachers want their students to develop accurate metacognitive judgments and ensure they acquire precise understanding of what they must still learn, for example, in self-regulated learning settings, the teacher should select specific tasks that give the learners substantial feedback about their actual knowledge state. When teachers aim to motivate their students and make them feel self-efficacious, for example, in a problem-based learning unit, they should use tasks that provide learners with less feedback.

Limitations

Order of posttest tasks

In our experimental design, we used the same order of posttest tasks for all participants; first free-recall posttest and then short-answer posttest. We chose this task sequence because free-recall and short-answer address different levels of memory accessibility. Free-recall tasks assess easily accessible knowledge. Additionally to high accessible knowledge, short-answer tasks also assesses low accessible knowledge by providing a specific cue for recall. This cue makes recall easier and allows us to assess harder-to-recall knowledge as well. We were interested in both types of accessibility. We stuck to the free-recall, then short-answer order of tasks to ensure that our short-answer tasks did not cue certain contents in the free-recall answers. We expected no carryover effects from the free-recall to short-answer tasks.

In this context, it is essential to distinguish the three cases. Firstly, a participant does not retrieve specific pieces of information in the free-recall task that were associated with the correct answer to a later short-answer task. In this case, there is no plausible carryover effect of activation. Secondly, a participant does retrieve a specific piece of information in the free-recall task that is the correct answer to a later short-answer task. As retrieving that piece of information is easier when responding to a short-answer task, that individual is more likely to retrieve the correct answer to the short-answer task with or without the preceding free-recall. As the free-recall pre-activates this piece of information in the

present case, that individual may have provided a faster answer to the short-answer task (which our study, however, did not measure). Thirdly, a participant retrieves specific pieces of information in the free-recall task that were associated with the correct answer to a later short-answer task. In this case, the activation of the associated knowledge in the free-recall task could have led to an activation of information corresponding to a subsequent correct answer via spreading activation. This activation of a correct answer later could have led to better recall while working on specific tasks (retrieval-induced facilitation effect, Chan 2006). However, studies directly addressing the issue of retrieval-induced facilitation without a delay (Chan 2010) found no evidence for a benefit of non-directly activated but associated material in an immediate posttest. The effect of retrieval-induced facilitation seems to develop only in conjunction with a delay. As we observed no delay in our posttest between the free-recall and short-answer tasks, there was very likely no effect of retrieving pieces of information in free-recall on retrieving associated information when answering short-answer tasks later.

Limitations of our findings

We confirmed most of our hypotheses. However, the holistic retrieval hypothesis was only confirmed with the free-recall posttest, but not with the short-answer posttest. There are several possible interpretations of these inconsistent findings. The first is that the free-recall posttest is more sensitive to differences because the short-answer tasks provide learners with retrieval cues. These retrieval cues may eliminate differences between conditions and could be responsible for the lack of evidence in the short-answer posttest. In this interpretation, the free-recall task would be the more informative one. A second interpretation is that the lack of evidence for the unspecific recall hypothesis is due to transfer-appropriate processing. As mentioned above, transfer-appropriate processing effects apply when different cognitive processes are involved (e.g. Veltre et al. 2016). Transfer-appropriate processing would assume that, in addition to the activation patterns, the relation between an intervention task and assessment task had an influence on our assessment. However, the interactions between the intervention and assessment tasks were not significant in any of our analyses in terms of direct effects. Hence, a transfer-appropriate processing interpretation is implausible in this case. We, therefore, reject this second alternative interpretation.

Further studies

Our findings suggest that a more detailed analysis of learning outcomes can establish a fruitful basis for drawing practical and theoretical implications. Our study's careful and detailed examination of learning outcomes has revealed evidence that is potentially relevant in further research.

Our study suggests that the task type selected for retrieval practice should depend on the nature of the learning content: a few central issues or many relevant information units. To further confirm this recommendation, this feature of learning content should be varied experimentally. For example, we could use learning material about the (three) different types in cognitive load theory (Sweller et al. 2011) and explain them via several examples. In such a case with three (few) central ideas, specific questions should be preferred. The second learning material could explain the cognitive load theory as a whole, with

important details and instructional design effects. In the latter case, unspecific free-recall might be preferable.

Another interesting direction in the current research of retrieval practice effects is the use of posttests given after a longer delay. In our study we provided the posttest after a one-week delay. In different studies (e.g. Rummer et al. 2017) retrieval practice interventions interact with the delays. Further studies could examine whether there are also interactions between delay and the effects of a task's specificity.

With respect to indirect effects: a limitation of this study is that we only assessed calibration, self-efficacy, or situational interest after retrieval practice. However, we did not investigate the influence of these variables on later remedial learning. Future studies should close this gap.

Conclusion

Overall, the type of recall task in retrieval practice makes a difference in learning. Teachers and instructional designers should be aware of task-specific *direct and indirect effects.* They should choose recall tasks that correspond to their main educational goals in a specific lesson. Matching the type of recall task to corresponding educational goals is necessary with respect to both learning outcomes (direct effects) as well as metacognition and motivation (indirect effects).

Funding Open Access funding enabled and organized by Projekt DEAL.

Compliance with ethical standards

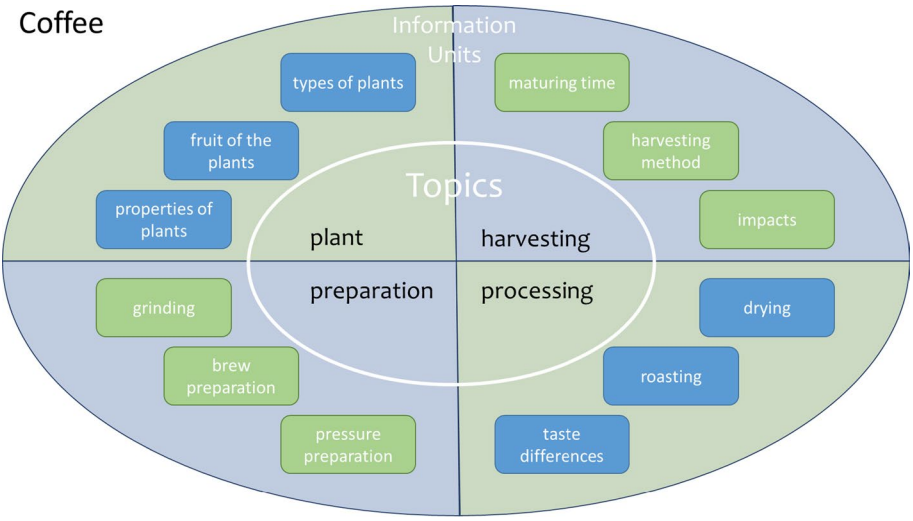
Conflict of interest The authors declare no conflicts of interest.

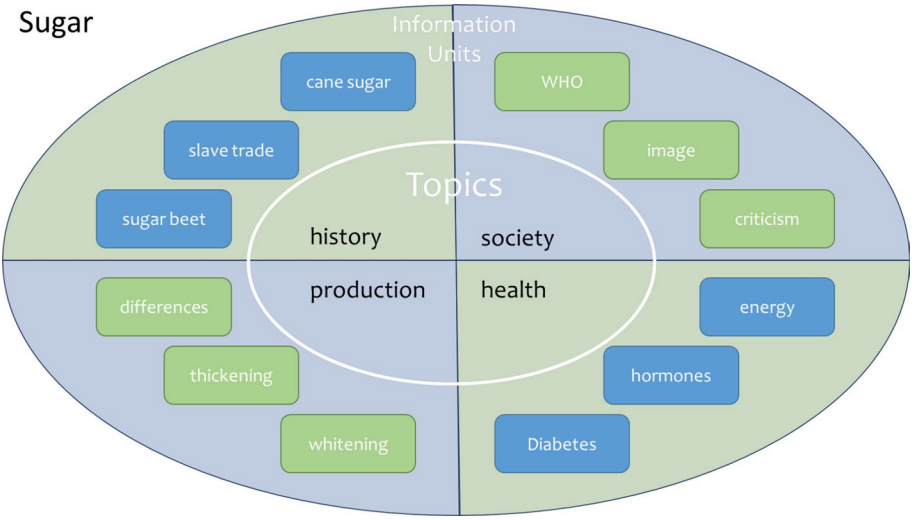
Ethical approval The whole experiment followed the rules set by the ethical guidelines of the German Psychological Society's (DGPs; 2004, CIII).

Informed consent Written informed consent was given by all participants.

Appendix

Appendix A Content Maps





Appendix B Material Piloting

Please rate on the following scales the text you just read.
How difficult was it to read the text?
In your opinion, was the text easy to understand?
Was the wording of the text appropriate?
Was the sentence length appropriate?
In your opinion, which sentences or paragraphs could have been improved?

Appendix C Texts

Coffee

Coffee consumption in Germany, Austria and Switzerland is constantly on the rise. This is easy to explain, because today coffee is not only a stimulating drink, but also a stimulant that can be consumed in style. This is evidenced by the increased interest in the origin of the beans, special varieties and the possibility of producing a good espresso at home.

The coffee plant is a tree and belongs to the group of the overgrowing, dicotyledonous, fused-crowned rubiaceae (reddish plants). It can reach a height of up to 15 m, but is cultivated in coffee plantations by permanent pruning as a shrub two to three meters high, which facilitates harvesting and increases yield. The coffee plant has large, green and shiny leaves that can grow up to 15 cm long. Only after three to five years do the plants produce the first white flowers, which sit on the long, tail-shaped branches. They bloom only for a few days. The fruits look like cherries, but their ripening period lasts up to ten months. They usually shine in a strong red, some varieties also in yellow. The coffee bean itself consists of several layers. Inside each of the coffee cherries there is a bean, which usually consists of two seeds, which are covered by a silver skin. In addition, there are the so-called pearl beans ("peaberry"), completely round seeds, as only one instead of one pair has developed in these cherries. Then follows the mucilage. It is 0.5–2 mm thick and consists mainly of sugar and pectin. Then follows the flesh of the fruit. When fresh, the color of the coffee beans is matt green in various nuances, depending on the origin and processing method.

The coffee plant is rather sensitive and does not cope well with extremes, which is why its cultivation is very time-consuming and the care accordingly time-consuming. Coffee plantations do not need much water, but regular irrigation and do not tolerate continuous rain. Since coffee belongs to the shade plants, a cultivation of the plant in the blazing sun would be advantageous for higher yields, because the plant gets under stress and produces more blossoms and more cherries; however, this kind of cultivation would leach out the soil more and require expensive fertilizers and pesticides. On some plantations, therefore, mixed cultivation is practiced: Other plants are placed between the coffee trees. These provide shade and shelter for birds, thereby reducing the need for pesticides and fertilizers.

The most famous of these are the plantation plants Arabica and Robusta. However, the seeds of at least five other species are also roasted and sold to make the drink coffee. Arabica plants are particularly sensitive and not frost-resistant. The temperature must not fall below zero degrees and the optimum temperature is 20 degrees. Furthermore, they are very susceptible to pests and fungi such as the coffee drill or the coffee rust: When in 1869 in

Sri Lanka the coffee plantations were attacked by it, it led to the birth of tea cultivation on the island. Arabica plants thrive at altitudes of 800 to 2200 m.

The Robusta on the other hand is much more resistant, but also has its peculiarities. Although it tolerates wetness and heat more easily, it is also more productive but has a longer ripening time and does not survive well in the cold at all. The temperature should be 5 degrees at least. In contrast to Arabica, it can also be cultivated in the lowlands up to a height of 800 m. In the beginning it was cultivated where no Arabica grew, but then its special qualities for espresso were discovered and efforts were made to improve the plant through crosses.

Long considered a cheap, inferior quality, today there are also very good Robusta qualities whose flavor can be described as nutty. Robusta is appreciated for its full body and rich crema in many Italian blends. In coffee blends for espresso, the Robusta is often used because of its celebratory and firmer cream, which complements the fine aromas of Arabica with body and fullness.

Coffee trees can live up to 50 years, but the yield decreases after about 20 years. As a rule, coffee is harvested once a year. It takes months for a ripe fruit to develop: Arabica between six and eight months, Robusta between nine and eleven months and Liberica or Excelsa even up to 14 months. This is why the latter two varieties are cultivated less frequently. In tropical regions it can sometimes even come to two flowers and thus to a main and secondary harvest. This is possible in Colombia or Kenya, for example.

The harvest time is determined by the degree of latitude and the cultivation height. As far as latitude is concerned, harvesting takes place between September and December north of the equator and between April and August south of the equator. At the equator itself, it would theoretically be possible to harvest all year round but here too there is at best a main and a secondary harvest. The cultivation height also has an influence on the duration of the ripening period and thus on the frequency of the harvest. At higher altitudes the ripening time of the coffee cherries rises while it decreases at lower altitudes.

The main harvesting period lasts between six to eight weeks but the entire harvesting period can extend to ten to twelve weeks, as not all fruits are ripe at the same time. The flowers develop at different times, which is why the cherries on the branches ripen differently at the same time. Over time, the cherries change color from green to yellow to red. Only when they are red do the cherries form mature acidity and full coffee flavor. Cherries in this state have a fruity, slightly sweet taste, which is reflected in the flesh of the bean. Only now are the important fats, oils and acids complete in the seed. These components are already essential for a good coffee taste in green coffee.

With selective picking only the ripe, red cherries are harvested by hand at 7–10-day intervals ("picking"). Each tree is visited three to four times, because the time from the first ripe, picked cherry to the last ripe cherry is up to 3 months. The amount of work involved is very high. The harvesting costs can increase threefold with this method, but the quality is usually better. In order to save time and money all cherries can also be stripped from the branches at the same time when most of them are ripe. This type of picking is called stripping. After stripping, the ripe fruits are mixed with unripe and overripe fruits, so they must be sorted afterwards. In some countries, such as Brazil, harvesting machines are already in use that work in a similar way to the stripping process. However, they are very expensive and can only be used in flatter areas. The cherries harvested in this way are placed in water for separation. Here the ripe and unripe cherries are separated: they float on the surface because they are lighter.

A picker needs 20 to 30 min per tree and harvests five to eight kilos of cherries, but after processing, drying and roasting only one to two kilos of roasted coffee remain. So, 100

kilos of cherries become twelve to 20 kilos of exportable coffee beans. The global average yield per hectare is around 600 kg of coffee. However, there is a wide range depending on the growing area. It depends, among other things, on the cultivation methods or the number of coffee plants.

There are several ways of preparing cherries which changes their taste characteristics. If the cherries are dried as a whole, they are called "naturals". If they are left to dry strongly and pulped in a raisin-like, just soft state, they are "pulped naturals". Ripe cherries that have been plucked, i.e. dried without the pulp but with the fruit mucilage of the mucilage, are called "semi-washed coffee". If the parchment skin is freed from the mucilage by wet fermentation, "washed coffee" is obtained; if this is done mechanically, the coffee is called "fully washed".

With every processing method, the coffee must dry to a moisture content of ten to twelve percent before packaging to ensure safe transport. Sometimes, additional hot-air-powered drying devices are also used for this purpose. The rule here is: the longer, the gentler.

The coffee beans are put in the hot-air dryer with about 20 percent moisture and in the warehouse with ten to twelve percent. The drier the coffee, the less risk there is during transport, but this results in a loss of weight and the associated loss of price. Around 1000 aroma substances are estimated to be found in every coffee bean: they are released by slow, fat-free heating—roasting. The specific character of a type of coffee is formed by heating. Depending on the variety or blend, there is an optimum roasting temperature and duration. However, not only the pure coffee is processed, but often a "blend", a mixture of two or more types of coffee, which can be strong or mild. The country of origin and the way the coffee is processed are also important. If you are used to a certain coffee, you want it to always taste almost the same, and this consistency can only be achieved with a blend. The art lies in putting together different types of coffee in such a way that their individual characteristics are advantageously combined.

The roasting process is also adapted to the subsequent processing. This is why there are different roasts for filter, mocha and espresso coffee, which are rarely interchangeable. If, for example, a typical roast for filter coffee is prepared as espresso, the result will probably taste unsatisfactory. Conversely, some fine aromas are lost when the coffee is roasted longer and darker, as the espresso is dominated by bittersweet nuances and roasted notes.

The moisture in the beans of the green coffee evaporates through the heat and thus increases the internal pressure. Evaporation and roasting gases such as carbon monoxide and carbon dioxide cause the bean to swell to twice its original size. However, the pressure can only escape when the cell walls of the bean have become brittle and crackle with a crackling noise. This crackling is called "first crack". The heat also causes a chemical reaction, the so-called Maillard reaction, in which carbohydrates such as sugar and free amino acids are broken down and converted into new compounds. This produces the roasted aromas and the bitter substances that are important for the coffee taste. The cracking is an important indicator for the roaster, since the bean loses more and more moisture from this tent point. For a light roast, the process is now interrupted. However, some varieties are roasted until a second crack and beyond.

Therefore, the roasting time is decisive for the aroma, but also for the acid formation of a coffee. The longer the coffee is roasted, the more complex the aroma compounds can be. If the beans are roasted very dark, more bitter substances are produced, which is why some varieties taste rather like bitter chocolate. For fruity, flowery nuances, the beans should remain lighter. This rule applies regardless of the type of preparation (espresso or filter coffee) for which the beans are intended.

The preparation of good coffee requires many aspects. For good coffee, the degree of grind and the freshness of the coffee grounds are essential prerequisites. To illustrate a drastic example: If hot water is poured over coffee beans, no coffee will be produced even after a long wait. To make coffee, you have to increase the surface area, in other words, the beans have to be crushed. The finer they are ground, the larger the surface and the better the hot water can extract the soluble substances.

During grinding, pressure is exerted on the bean until it breaks. The aim is to achieve as uniform a particle size as possible. If the knives are too blunt, they squeeze these particles into larger crumbs, allowing the water to flow through irregularly and faster during extraction. Evenly ground and compressed coffee particles provide the water with better resistance so that it stays longer in the coffee and can extract more content. However, this does not mean that the finer the better. If the ground coffee has been ground too finely, substances are extracted during preparation that we actually do not want to taste, such as bitter substances. The ideal degree of grinding depends on the result of the extraction: If the water flows too slowly, a coarser setting must be selected; if it flows too quickly, it is correspondingly finer. The finer the coffee is ground, the longer the grinding process takes. For coffee preparation by brewing, coarse to medium coarse grinding is possible depending on the machine. For filter coffee a medium fine setting is suitable, for espresso a very fine setting and for Turkish coffee powder it must be finely ground.

There are many methods of making good coffee. There is actually no objective evaluation here, but only personal charts, because the simple brew infusion has just as many qualities as the ambitious espresso. The oldest method of making coffee is probably the brew infusion. From the holidays in Greece, Turkey or in the Arabic area one knows the oriental method and/or the Turkish or Greek coffee. The typical small metal pot is filled with water. Then you add finely ground coffee powder and sugar if desired and boil this mixture up. Especially in the Arab world, spices such as cardamom, cloves or saffron are sometimes also cooked. When the foam rises, take the pot off the fire, wait briefly and let it boil a second time, sometimes even a third time. It is served in small mocha cups.

Another brewing method is the filter method. The paper filter, which was patented by the German housewife Melitta Bentz in 1908, was groundbreaking for making coffee at home. The medium-fine ground coffee grounds are filled into the filter, then evenly infused with a hot surge of non-boiling water so that the coffee is moistened and can swell. Then add the rest of the water. The liquid should drain off evenly and leave a flat coffee bed. Flow-through brewers are also used, but the extraction time cannot be adjusted. The latest models in this field are based on the "on demand" principle: the coffee is freshly ground for each cup of filter coffee.

The espresso is the opposite. During espresso preparation, the hot water is pressed through the coffee grounds at high pressure. The question arises as to how much bar pressure one can call the extract "espresso". The espresso lover would say: no espresso without crema. There are many different techniques for preparing espresso which differ primarily in how the pressure for the espresso is produced. In the meantime, most people are using electric pumps to generate the desired water pressure of 9 bars. The first important factor when preparing an espresso is the so-called rope. This term describes the compression of the coffee grounds in the brewing group of the espresso machine. The filled coffee grounds are compressed with a round pressure on a metal disk, so that water can flow evenly through all places of the coffee cake. Then the pre-brewing takes place, in which hot water is added to the brewing group at low pressure. A valve is then opened, which presses water at 9 bars onto the coffee grounds and brews the coffee. This valve delivers water at a constant pressure of 9 bar until the barista stops the process. An ideal espresso is produced when an

espresso with 25 ml has been produced after 25 s. If the water runs faster through the flour, the complete aroma is usually not extracted, and the coffee is too tasteless.

Sugar

The business of sugar. It is white, sweet and very popular among people: every German alone eats around 35 kg of sugar a year. A daily dose of about 100 g is four times as much as experts allow us to eat. The confectionery industry in particular is doing good business with this desire. After all, it achieved sales of around 13 billion euros in Germany in 2009. The appetite for sugar has increased significantly in Germany: According to the German Nutrition Society (Deutsche Gesellschaft für Ernährung), consumption per capita has increased by 400 g every year since 1995. This is mainly due to the fact that Germans are eating more and more products containing sugar. Doctors and consumer protectors have been warning about this development for a long time, as they see increased sugar consumption as one reason why people in Europe are getting fatter and fatter. The sugar and confectionery industries take a different view. And that is why they have a special interest in sugar maintaining a positive image in the public eye. Under no circumstances should sugar be associated with obesity or tooth decay. For example, they advertise products containing sugar, such as Nutella and Coca Cola, which are associated with sporting activity and joie de vivre. A new advertising strategy of well-known companies is to give the brand a new coat of paint. For some time now, Coca Cola, Pepsi and Co. have been experimenting with natural sweeteners such as Stevia rebaudiana. Stevia is 300 to 400 times sweeter than sugar and yet extremely low in calories. Stevia is supposed to make Coca Cola healthier and greener. “Coca Cola life”, a mixture of sugar and stevia, is already sold in Argentina, Mexico and Chile.

“Open your naturalness” says an advertising slogan, the first sip will change everything. On the website you can read that “Coca Cola life” even leads to a happier and healthier life. A look at the list of ingredients shows, however, that even the green varieties of Pepsi and Cola are only minimally healthier or better: less harmful to health. The new products contain just one third less sugar than the originals. The bottle versions are fully recyclable and consist of 30 percent vegetable material. The remaining 70 percent, however, consists of fossil raw materials. Commercially available PET bottles are at least as environmentally friendly. So above all, the image becomes greener.

The World Health Organization (WHO) introduced a much stricter guideline for the sugar content in food in 2015. Instead of the current maximum of ten percent, the proportion of hidden sugar should now be as low as five percent. The new recommendation deals with types of sugar such as sucrose, fructose and glucose that manufacturers, cooks or consumers add to their food and drinks. Sweet whey powder, dextrose, glucose syrup, maltodextrin and whey product—are only a small selection of products behind which sugar is ultimately found. If one of these products is found on the list of ingredients of convenience foods, the known sugar content slips further back. This gives the impression that a food contains less sugar and is healthier. However, the natural sugar content in fruit, vegetables and milk is irrelevant, since according to the WHO there is no indication that it is harmful to health. The limit, halved to five percent, means that an adult should not consume more than 25 g of sugar per day, i.e. six teaspoons of it. A can of lemonade alone contains an average of ten teaspoons of sugar.

The WHO pointed out that the sugar content of industrial foods is often difficult to detect. One tablespoon of ketchup, for example, contains a whole teaspoon of sugar. The

WHO therefore calls for better labelling of the hidden sugar content in food and drink and for a restriction on advertising of sugary foods to children. The organization also advocates a dialogue with the food industry to reduce the sugar content of its products.

The food industry in Germany criticized the new WHO guideline as a "false solution" with dubious scientific justification. Obesity is not caused by diet alone, but also by lack of exercise, hereditary and socio-economic factors as well as stress, explained the Bund für Lebensmittelrecht und Lebensmittelkunde in Berlin.

The triumph of the white crystals—sugar—has conquered and changed the world. Nowadays sugar is one of the most common products, but there was a time when this was not the case. 500 years ago, there was virtually no sugar on the market, especially in Europe. A few sacks, among other spices, were laboriously transported from faraway Asia to Europe, and a few small sugar cane plantations existed around the Mediterranean. In the South Pacific region, sugar cane has been known for several thousand years. But its cultivation is laborious and so there was no abundance of sugar there either. The conquest of the world by sugar began with the conquest of the New World by Christopher Columbus.

Because of the tropically warm climate, it soon became clear to the conquerors that sugar cane could be grown here. On board one of the first ships they therefore took some sugar cane seedlings with them to the Caribbean. Although the plants grew and prospered, the work in the fields was extremely strenuous. The locals refused to help and paid for this attitude with their lives. Only 50 years later all Native Americans had been wiped out. Now the conquerors had to come up with something so that the plants would not die again. Since the few small sugar cane plantations in the Mediterranean region were already managed by slaves—the plague in Europe had decimated the regular workforce—the solution for sugar cane in the Caribbean was obvious. The Spanish bought slaves from Africa.

In the following 400 years more than 10 million Africans were shipped to America. For one to two months the 400 to 600 slaves on board of the ships had to lived chained in a crouched or lying position. Only strong and healthy people could endure this. More than a third did not survive the crossing. Arriving overseas, they had to cope with the extremely hard work on the sugar cane fields every day from morning to evening. But although many slaves did not live long here either, the business was worthwhile for the plantation owners: with an annual output of three to four tons of sugar, a slave had “amortized” itself after two years purely mathematically. A triangular trade dominated trade on the Atlantic for four centuries: sugar came from America to Europe and weapons, spirits and cotton fabrics from here to Africa and people from Africa to America.

Frederick the Great of Prussia was annoyed in the middle of the eighteenth century about the high sugar prices and his dependence on the Spaniards and Englishmen, who were meanwhile also active in the sugar cane business. He commissioned the Berlin chemist Andreas Sigismund Marggraf to look for a native plant that also contained a larger quantity of sugar. And Marggraf found what he was looking for in 1747. He isolated the sweet crystals from the beet, which until then had been considered a poor man’s food. But it was only his successor Franz Carl Achard who managed to grow the beet in such a way that the sugar yield really paid off. And it was not until 1802 that Achard was able to open the world’s first beet sugar factory in Kunern, Silesia. This was followed by the first boom in the European beet sugar industry thanks to Napoleon Bonaparte, the emperor of the French. With his Continental Blockade in 1806, he prevented the import of Caribbean cane sugar and temporarily turned beet growers into monopolists. But after the end of the blockade, cane sugar flooded the market and all but one French sugar factory had to close. This factory, however, was the nucleus of the European beet sugar industry, which from now on gained more and more ground.

Before the production of sugar was known, sugar cane was chewed and sucked to obtain its sweet taste. In the tropical countries where sugar cane is grown, you can still find fresh cane on the regional markets today, which is used as it was originally to extract the sweet, juicy pulp or to squeeze a kind of soft drink out of it. Nowadays, sugar cane is cultivated worldwide mainly in South America, Asia, Africa and Australia. The well-known cane sugar is primarily used for sweetening, for example in confectionery and bakery products. The golden brown sugar is also often used for decorating. Cane sugar also serves as the basis for one of the most popular cocktails in Germany, the caipirinhia. Many hundreds of years ago, sugar was still pure luxury and was only used sparingly on special occasions. Cane sugar is an essential part of everyday life and is part of the lifestyle. Brazil also uses the sugar cane plant to produce alcohol fuel, which is much cheaper than conventional petroleum-based fuels. 717.

The well-known white sugar, as we know it from the supermarket, can be made from beets or sugar cane. It is refined many times until the molasses are completely removed.

Clumsy and not sweet at all—that's the impression you could get when you look at a sugar beet. But inside it hides a sweet secret: sugar. The sugar beets planted today contain about 20 percent sugar: sucrose. Today, the inconspicuous beets from Germany cover almost our entire sugar requirement. For this purpose, about 4 million tons of beets are harvested every year. But it's a long way from the field to the shelf, taking the beet through highly technical and complicated factory facilities. In autumn, large machines harvest the beet fields, which are spread over about 400,000 hectares throughout Germany. Only three large companies share the sugar market in Germany. The farmers deliver the raw material beet to the sugar factories on time and only on order. Because the delivery of the many beets to a few factories is not only a logistical challenge. Only when the beet is processed quickly after harvesting does it degrade little of its valuable ingredient, sugar. It only takes about ten hours from delivery to the factory to the finished sugar.

The process continues with a bath for the beets. A special beet washing machine and large quantities of water clean the beets of soil and dirt. Afterwards it goes over treadmills into the cutting machine: From it the so-called beet chips come, from which the sugar can be more simply removed. In a tower about 20 m high, the beet pulp travels from bottom to top, while hot steam is conducted over it. The sugar escapes from the plant cells and dissolves in the water.

The sugar cane must also be processed as soon as possible after harvesting, as it is quickly attacked by microorganisms that break down the sugar. In contrast to beet, the tube is ground into small pieces in a large mill. The juice is squeezed out of these pieces between rollers. Considerable quantities of fiber remain—the bagasse, which is returned to the field as a natural fertilizer.

The resulting sugar solution, the raw juice from beet or cane, is heated in large boilers. However, the solution still contains unwanted impurities. These are removed by so-called carbonation with chemicals until a light yellow, clear juice remains. To increase the sugar content, the raw juice is thickened in large boilers—until a sugar content of about 70 percent is reached. But the whole thing has little to do with household sugar: it is a viscous brown mass. One important step is still missing: the sugar in the juice must crystallize. At low pressure and around 70 degrees Celsius, sugar crystals of pure household sugar (sucrose) slowly form in a kettle. If they are large enough, they are separated from the remaining juice.

A centrifuge separates the crystals from the brown liquid, molasses: the centrifugal force presses them against a sieve through which only the liquid molasses can flow. The remaining sugar slowly changes color from dark brown to white. This sugar quality—called

"affinade"—is already sufficient for the food industry. But in order to produce particularly pure granulated sugar—the "refined sugar"—the white sugar is dissolved and crystallized again. Only then is it pure enough to end up on the supermarket shelf as household sugar.

An adult human being consists of about one percent of different sugar molecules. These carbohydrates are part of many cells and tissues. But above all, sugar is an important source of energy. The brain even derives its energy exclusively from sugar, which is constantly provided by the blood. Strictly speaking, blood sugar is glucose (dextrose). Glucose is absorbed from food and is also produced by the organism itself. The energy carrier glucose reaches all organs and tissues of the body via the blood, is absorbed by the cells and converted into energy. Glucose is completely degraded to carbon dioxide and water in the cytoplasm and mitochondria. At the end of the degradation process, the cell uses the released energy to obtain the energy-rich compound ATP, which is required for many metabolic processes as a universal source of energy for the organism. A control circuit ensures that the concentration of blood sugar remains constant and that the right amount of "fuel" is always available.

The most important hormone in the sugar balance is insulin. It is necessary for the cells of the muscles and fat tissue to be able to absorb glucose molecules from the blood. Like a key, the hormone opens channels in the cell membranes through which glucose then enters the cell interior. Insulin thus ensures that the cells have enough energy at their disposal as well as a constant blood sugar level. Insulin is produced in the pancreas. Specialized cells, which are distributed like islets in the metabolic organ ("islet cells"), produce the hormone as required and release it into the blood. If the pancreas produces too little insulin, glucose can only enter the cells to a reduced extent and the blood sugar level rises.

The brain cannot obtain energy from fat—it is dependent on the sugar in the blood. That is why the body prevents this: If the blood sugar drops too quickly, it immediately throttles insulin production. In addition, it releases another hormone: glucagon, the antagonist of insulin. The glucagon ensures that new sugars are formed in the liver from protein building blocks. It also releases previously stored sugar (glycogen) from the muscles. Together, insulin and glucagon ensure that the blood sugar in healthy people is constantly kept between 80 and 180 mg of glucose per 100 ml of blood.

If the glucose concentration in the blood is too high, there is a diabetes. The medical term for this is "diabetes mellitus". Translated this means approximately "honey-sweet flow". And in fact, in diabetes the flowing blood is sweet and often the urine too. The disease is very individual and there are two completely independent ways in which it develops. That is why doctors differentiate between types 1 and 2 in diabetes.

Diabetes mellitus type 1 used to be called adolescent diabetes because this form of diabetes typically begins in childhood and adolescence. In type 1 diabetics, the body's own immune system is directed against the insulin-producing cells of the pancreas and destroys them. Insulin production is therefore reduced and often comes to a complete standstill. As a result, the blood sugar level rises. In type 1 diabetes, the body's own insulin deficiency must always be compensated by the administration of insulin as a drug. When the immune system is directed against the invaders, it unintentionally destroys the important insulin cells. Doctors call this phenomenon autoimmunity.

The vast majority (95 percent) of diabetics are type 2 diabetics. This form was formerly also called adult onset diabetes, as it almost exclusively affected older people. However, young people and even children are increasingly developing this form of diabetes, the causes of which are fundamentally different from those of type 1: Type 2 diabetes is mainly caused by obesity and lack of exercise. But even in this form of diabetes, the hereditary factors form the basis on which the disease of sugar metabolism can develop. Parents probably

inherit how well the specialized islet cells of the pancreas react to an increased insulin requirement. If insulin production is limited due to genes, an increased insulin requirement cannot be met and diabetes develops. And many people have an increased insulin requirement—especially those who are overweight or lack exercise. The reason for this is that the cells of fat and muscle tissue react weaker to the metabolic hormone. Doctors call this insulin resistance.

Appendix D Questions

Coffee

Plant

Which types and characteristics of the "coffee cherry" do you know?
Which requirements of the coffee plant must be considered during cultivation?
What types of coffee plants do you know of and how do they differ?

Harvesting

Which factors does the maturing time of the coffee beans depend on?
Why are there different ripening beans on a tree and what effects does this have on the taste of the coffee been?
What harvesting methods do you know and how do they differ?

Processing

Which drying methods can you name and how do they differ?
Which reactions take place when roasting coffee?
Why is coffee roasted and on which factors is the roasting dependent

Preparation

Why is coffee ground and which characteristics of grinding are relevant for the taste?
Which brewing methods can you name and how do they differ?
Which factors are relevant for the preparation of a good espresso?

Sugar

Society

How has sugar consumption developed over the years?
What image do well-known sugar companies convey and what measures are being taken to achieve this?
What did the WHO currently draw attention to and what consequences did this have?

History

How and under what conditions did sugar cane come to Europe?
What was the history of the beet?
Where and how is sugar cane still used today?

Production

Describe the first processing step after harvesting the beet.
How is the sugar solution from the raw juice further processed?
What steps are necessary to make sugar white and describe them.

Health

What role does sugar play in the human body?
Describe how the two essential hormones in the sugar balance are related to each other and which functions they fulfil.
What is the essential difference between the two forms of diabetes?

References

Adesope, O. O., Trevisan, D. A., & Sundararajan, N. (2017). Rethinking the Use of Tests : A Meta-Analysis of Practice Testing. *Review of Educational Research, 87*(3), 659–701. https://doi.org/10.3102/0034654316689306.

Alexander, P. A. (2013). Calibration: What is it and why it matters? An introduction to the special issue on calibrating calibration. *Learning and Instruction, 24*(1), 1–3. https://doi.org/10.1016/j.learninstruc.2012.10.003.

Anderson, J. R., & Reder, L. M. (1999). The fan effect: New results and new theories. *Journal of Experimental Psychology: General, 128*(2), 186–197. https://doi.org/10.1037/0096-3445.128.2.186.

Arnold, K. M., & McDermott, K. B. (2013). Test-potentiated learning: distinguishing between direct and indirect effects of tests. *Journal of Experimental Psychology. Learning, Memory, and Cognition, 39*(3), 940–945. https://doi.org/10.1037/a0029199.

Bandura, A. (1997). *Self-efficacy: The exercise of control*. New York: Freeman.

Bandura, A. (2006). Guide for constructing self-efficacy scales. *Self-Efficacy Beliefs of Adolescents*. https://doi.org/10.1017/CBO9781107415324.004.

Butler, A. C., & Roediger, H. L. I. I. I. (2007). Testing improves long-term retention in a simulated classroom setting. *European Journal of Cognitive Psychology, 19*(4–5), 514–527. https://doi.org/10.1080/09541440701326097.

Carpenter, S. K. (2012). Testing enhances the transfer of learning. *Current Directions in Psychological Science, 21*(5), 279–283. https://doi.org/10.1177/0963721412452728.

Carpenter, S. K., Lund, T. J. S., Coffman, C. R., Armstrong, P. I., Lamm, M. H., & Reason, R. D. (2016). A classroom study on the relationship between student achievement and retrieval-enhanced learning. *Educational Psychology Review, 28*(2), 353–375. https://doi.org/10.1007/s10648-015-9311-9.

Carpenter, S. K. (2009). Cue strength as a moderator of the testing effect: the benefits of elaborative retrieval. *Journal of Experimental Psychology. Learning, Memory, and Cognition, 35*(6), 1563–1569. https://doi.org/10.1037/a0017021.

Chan, J. C. K. (2010). Long-term effects of testing on the recall of nontested materials. *Memory, 18*(1), 49–57. https://doi.org/10.1080/09658210903405737.

Chan, J. C. K., & McDermott, K. B. (2007). The testing effect in recognition memory a dual process account. *Journal of Experimental Psychology. Learning, Memory, and Cognition, 33*(2), 431–437. https://doi.org/10.1037/0278-7393.33.2.431.

Chan, J. C. K., McDermott, K. B., & Roediger, H. L. (2006). Retrieval-induced facilitation: initially nontested material can benefit from prior testing of related material. *Journal of Experimental Psychology. General, 135*(4), 553–571. https://doi.org/10.1037/0096-3445.135.4.553.

Davis, D., Chen, G., Zee, T. Van Der, Hauff, C., & Houben, G. (2016) Retrieval Practice and Study Planning in MOOCs Exploring Classroom-Based Self-Regulated Learning Strategies at Scale. *Proceedings of the 11th European Conference on Technology Enhanced Learning (EC-TEL)*. https://doi.org/10.1007/978-3-319-45153-4_5.

Duchastel, P., & Nungester, R. (1982). Testing effects measured with alternate test forms. *The Journal of Educational Research*.

Dunlosky, J., Rawsona, K., Marsh, E. J., Nathan, M. J., & Willingham, D. T. (2013). Improving students' learning with effective learning techniques: promising directions from cognitive and educational psychology. *Psychological Science in the Public Interest, 14*(1), 4–58. https://doi.org/10.1177/1529100612453266.

Endres, T., & Renkl, A. (2015). Mechanisms behind the testing effect: An empirical investigation of retrieval practice in meaningful learning. *Frontiers in Psychology*, *6*, 1054. https://doi.org/10.3389/fpsyg.2015.01054.

Faul, F., Erdfelder, E., Lang, A.-G., & Buchner, A. (2007). G*Power: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods, 39*(2), 175–191. https://doi.org/10.3758/BF03193146.

Glover, J. (1989). The "testing" phenomenon: Not gone but nearly forgotten. *Journal of Educational Psychology, 81*(3), 392–399. https://doi.org/10.1037//0022-0663.81.3.392

Hidi, S., & Renninger, K. A. (2006). The four-phase model of interest development. *Educational Psychologist*. https://doi.org/10.1207/s15326985ep4102_4.

Kang, S. H. K., McDermott, K. B., & Roediger, H. L. (2007). Test format and corrective feedback modify the effect of testing on long-term retention. *European Journal of Cognitive Psychology, 19*(4–5), 528–558. https://doi.org/10.1080/09541440601056620.

Karpicke, J. D., & Aue, W. R. (2015). The testing effect is alive and well with complex materials. *Educational Psychology Review, 27*(2), 317–326. https://doi.org/10.1007/s10648-015-9309-3.

Karpicke, J. D., Lehman, M., & Aue, W. R. (2014). *Retrieval-Based Learning : An Episodic Context Account* (Vol. 61).

Koriat, A. (2012). The relationships between monitoring, regulation and performance. *Learning and Instruction, 22*(4), 296–298. https://doi.org/10.1016/j.learninstruc.2012.01.002.

Lehman, M., Smith, M. A., & Karpicke, J. D. (2014). Journal of Experimental Psychology : Learning , Memory , and Cognition Toward an Episodic Context Account of Retrieval-Based Learning : Dissociating Retrieval Practice and Elaboration Toward an Episodic Context Account of Retrieval-Based Learning : Dissoc.

Linnenbrink-Garcia, L., Patall, E. A., & Messersmith, E. E. (2013). Antecedents and consequences of situational interest. *British Journal of Educational Psychology, 83*(4), 591–614. https://doi.org/10.1111/j.2044-8279.2012.02080.x.

Montoya, A. K., & Hayes, A. F. (2017). Two-condition within-participant statistical mediation analysis: A path-analytic framework. *Psychological Methods, 22*(1), 6–27. https://doi.org/10.1037/met0000086.

Morris, C. D., Bransford, J. D., & Franks, J. J. (1977). Levels of processing versus transfer appropriate processing. *Journal of Verbal Learning and Verbal Behavior*. https://doi.org/10.1016/S0022-5371(77)80016-9.

Nelson, T. O., & Narens, L. (1990). Metamemory: A Theoretical Framework and New Findings. *The Psychology of Learning and Motivation, 26,* 125–173. https://doi.org/10.1016/S0079-7421(08)60053-5.

Nückles, M., Hübner, S., & Renkl, A. (2009). Enhancing self-regulated learning by writing learning protocols. *Learning and Instruction, 19*(3), 259–271. https://doi.org/10.1016/j.learninstruc.2008.05.002.

Pan, S. C., & Rickard, T. C. (2018) Annual Meeting of the Psychonomic Society in Vancouver, BC This research was supported by an American Psychological Association (APA) Early Graduate Student Researcher Award and a National Science. https://doi.org/https://doi.org/10.1037/bul0000151.

Pass, F. G. W. (1992). Training Strategies for Attaining Transfer of Problem-Solving Skill in Statistics - a Cognitive-Load Approach. *Journal of Educational Psychology, 84*(4), 429–434. https://doi.org/10.1037/0022-0663.84.4.429.

Pyc, M. A., Agarwal, P. K., & Roediger, H. L. III. (2014). Test-enhanced learning. In V. A. Benassi, C. E. Overson, & C. M. Hakala (Eds.), *Applying science of learning in education: Infusing psychological science into the curriculum* (p. 78–90). Society for the Teaching of Psychology.Pyc, M. a., & Rawson, K. a. (2009). Testing the retrieval effort hypothesis: Does greater difficulty correctly

recalling information lead to higher levels of memory? *Journal of Memory and Language*, *60*(4), 437–447. https://doi.org/10.1016/j.jml.2009.01.004.
Pyc, M. A., & Rawson, K. A. (2010). Why testing improves memory: Mediator effectiveness hypothesis. *Science, 330*(6002), 335. https://doi.org/10.1126/science.1191465.
Rawson, K. A., & Zamary, A. (2019). Why is free recall practice more effective than recognition practice for enhancing memory? Evaluating the relational processing hypothesis. *Journal of Memory and Language, 105,* 141–152. https://doi.org/10.1016/j.jml.2019.01.002.
Roediger, H. L., Putnam, A. L., & Smith, M. A. (2011). Ten Benefits of Testing and Their Applications to Educational Practice. *Psychology of Learning and Motivation, 55,* 1–36. https://doi.org/10.1016/B978-0-12-387691-1.00001-6.
Rotgans, J. I., & Schmidt, H. G. (2011). Situational interest and academic achievement in the active-learning classroom. *Learning and Instruction, 21*(1), 58–67. https://doi.org/10.1016/j.learninstruc.2009.11.001.
Rotgans, J. I., & Schmidt, H. G. (2014). Situational interest and learning: Thirst for knowledge. *Learning and Instruction, 32,* 37–50. https://doi.org/10.1016/j.learninstruc.2014.01.002.
Rotgans, J. I., & Schmidt, H. G. (2017). The relation between individual interest and knowledge acquisition. *British Educational Research Journal, 43*(2), 350–371. https://doi.org/10.1002/berj.3268.
Rouder, J. N., Morey, R. D., Speckman, P. L., & Province, J. M. (2012). Default Bayes factors for ANOVA designs. *Journal of Mathematical Psychology, 56*(5), 356–374. https://doi.org/10.1016/j.jmp.2012.08.001.
Rowland, C. A. (2014). The effect of testing versus restudy on retention: A meta-analytic review of the testing effect. *Psychological Bulletin, 140*(6), 1432–1463. https://doi.org/10.1037/a0037559.
Rummer, R., Schweppe, J., Gerst, K., & Wagner, S. (2017). Is testing a more effective learning strategy than note-taking? *Journal of Experimental Psychology: Applied*, *23*(3), 293–300.
Ryan, R. M., & Deci, E. L. (2000). Self-determination theory and the facilitation of intrinsic motivation, social development, and well-being. *The American Psychologist, 55*(1), 68–78. https://doi.org/10.1037/0003-066X.55.1.68.
Schiefele, U. (1990). Thematisches Interesse, Variablen des Lernprozesses und Textverstehen. *Zeitschrift Fuer Experimentelle Und Angewandte Psychologie, 37*(2), 304–332.
Schmitt, N. (1996). Uses and abuses of coefficient alpha. *Psychological Assessment, 8*(4), 350–353. https://doi.org/10.1037/1040-3590.8.4.350.
Schraw, G. (2009). A conceptual analysis of five measures of metacognitive monitoring. *Metacognition and Learning, 4*(1), 33–45. https://doi.org/10.1007/s11409-008-9031-3.
Schunk, D. H. (1985). Self-efficacy and classroom learning. *Psychology in the Schools, 22*(2), 208–223. https://doi.org/10.1002/1520-6807(198504)22:2%3c208::AID-PITS2310220215%3e3.0.CO;2-7.
Smith, M. A., Blunt, J. R., Whiffen, J. W., & Karpicke, J. D. (2016). Does Providing Prompts During Retrieval Practice Improve Learning? *Applied Cognitive Psychology, 30*(4), 544–553. https://doi.org/10.1002/acp.3227.
Smith, M. A., & Karpicke, J. D. (2014). Retrieval practice with short-answer, multiple-choice, and hybrid tests. *Memory, 22*(7), 784–802. https://doi.org/10.1080/09658211.2013.831454.
Sweller, J., Ayres, P., & Kalyuga, S. (2011). *Cognitive load theory*. Heidelberg: Springer.
Vansteenkiste, M., Simons, J., Lens, W., Sheldon, K. M., & Deci, E. L. (2004). Motivating learning, performance, and persistence: The synergistic effects of intrinsic goal contents and autonomy-supportive contexts. *Journal of Personality and Social Psychology, 87*(2), 246–260. https://doi.org/10.1037/0022-3514.87.2.246.
Veltre, M. T., Cho, K. W., Neely, J. H., Veltre, M. T., Cho, K. W., Transfer-, J. H. N., & Neely, J. H. (2016). Transfer-appropriate processing in the testing effect Transfer-appropriate processing in the testing effect. *Memory, 8211*(8), 1229–1237. https://doi.org/10.1080/09658211.2014.970196.
Wetzels, R., Matzke, D., Lee, M. D., Rouder, J. N., Iverson, G. J., & Wagenmakers, E. J. (2011). Statistical evidence in experimental psychology: An empirical comparison using 855 t tests. *Perspectives on Psychological Science, 6*(3), 291–298. https://doi.org/10.1177/1745691611406923.
Zaromb, F. M., & Roediger, H. L. (2010). The testing effect in free recall is associated with enhanced organizational processes. *Memory & Cognition, 38*(8), 995–1008. https://doi.org/10.3758/MC.38.8.995.